A BAYESIAN APPROACH FOR BINARY OUTCOME A/B TESTS

Ramesh Subramonian, Ranjeet Tate, Michael Shire, Abhi Singh

1 Introduction

In a *binary* experiment, each trial has only two possible outcomes, a *success* (1) or a *failure* (0).

A population being experimented on has an *intrinsic* property —for a binary outcome, the true probability of success— and a test is intended to measure this in order that (business) decisions can be made based on its value.

In an A/B test there are two populations or branches, each of which is exposed to a different treatment. One branch is called the *control*, denoted in this article by B, which is usually a standard or baseline, and another is called the *variant*, denoted by A. The treatments could be different advertising campaigns, different form or content of a landing page on a website or different product offerings. The populations have *intrinsic success probabilities* p_A and p_B. The *goal of the analysis is to provide the client with a recommendation about choosing A or B*[1] based on identified business needs, using the data and justified by statistics.

Of course, in a finite test, we cannot know the intrinsic success probabilities with certainty. Consider a single population: A given test consists of trials conducted on a finite sample of the population, and the number of trials n and successes m are recorded. From this behavioral data, the success probability p of the population is to be probabilistically inferred, e.g., 60 successes out of a 100 trials is most *likely* to have arisen from a population with success probability $p = 0.6$, but it could have arisen, albeit with less likelihood, from a population with $p = 0.5$ or $p = 0.7$, or indeed any value in $(0.0, 1.0)$. Hence, from the data n, m, we infer the *likelihood* of p, which is a probability distribution on $\{p\}$.

In an A/B test on two populations, the experiment will yield a count of the trials and successes in each branch, (n_A, m_A) and (n_B, m_B). From this data we infer the likelihood on *2-dimensional probability space* $\{(p_A, p_B)\}$.

There are different approaches to what we infer and how we infer it, and we prefer the Bayesian to the Frequentist[2] for reasons outlined in the Appendix (Section **??**) and discussed in more detail in a later article[3]. In Section 3.1 we describe the Bayesian Approach and use it to construct the

[1] Paraphrasing from Klugman *et al* "Loss Models", 2nd Ed. Wiley (2004), pg. 419: "...the process must end with a winner. While qualifications, caveats etc. are often necessary, a commitment is required.".

[2] See `http://jakevdp.github.io/blog/2014/03/11/frequentism-and-bayesianism-a-practical-` for a good exposition of the two approaches and the calculational and interpretational differences.

[3] See `frequentist_vs_bayesian.pdf`.

likelihood of a single population's success probability p. Then in Section 3.3 we construct the joint likelihood for the intrinsic success probabilities p_A and p_B of the two populations based on the experimental data. We use the data from the experiment to construct the posterior probability distribution (or *likelihood*) of (p_A, p_B). Fig. 1 is the two dimensional space of intrinsic probabilities $\{(p_A, p_B)\}$, on which we've shown a contour plot of the likelihood function, see the caption for additional explanation.

Figure 1: Contour Plot of the Likelihood of (p_A, p_B) for $(n_A, m_A, n_B, m_B) = (12, 10, 12, 7)$. The contours correspond to points of equal likelihood. The insides of high-value contour lines represent regions of high likelihood, with the peak at $(10/12, 7/12)$.

The implicit hypothesis that underlies most A/B experiments is that "A is better than B". For testing this hypothesis it is enough to consider the *difference* between success probabilities —i.e. $p_A - p_B > 0$[4]. (Before delving into the issue of statistics, assume for a moment that we know both p_A and p_B with certainty.) However, in most situations it is not enough that A be simply better than B: Siroker and Koomen themselves include an example of a test comparing a page with a static ad to one with a video[5]. The variant with the video was better in terms of both click-through and conversion rates, but the costs of producing the video and displaying it were deemed too high to launch the video version at scale. So while the difference was statistically significant it wasn't *important* enough from a business perspective. A complete approach requires the analyst or business client to establish an *acceptance level* for the metric M_{Acc} —e.g. $p_A - p_B > 0.1$ for the difference between A and B— which would lead to a recommendation of A over B. But now note that the analyst faces a choice of *metric* by which to compare p_A and p_B. E.g., in order to trigger action, do we want the *difference* between success probabilities to exceed 0.1 —i.e. $p_A - p_B > 0.1$— or do we want the success probability to *lift* by 20% —i.e. $p_A > 1.2 * p_B$? There are an immeasureable number of possible metrics and one has to determine a metric that correlates linearly with business goals[6].

We consider three metrics, discussed in Section 4. The first two (the probability difference and lift mentioned above) are very commonly used for Binary A/B testing but suffer some problems which we point out at the end of Section 4.2. In Section 4.3 we present the *odds factor*, which —if the probability being measured is the *page transition probability*, that of moving to any other page on the website— is the average number of pages visited. Since number of page visits correlates linearly to the number of conversion opportunities and hence revenue, this is a good metric from a business perspective.

[4] See e.g. *A/B Testing*, Dan Siroker and Pete Koomen, Wiley (2013)

[5] *ibid.* "Fail Fast and Learn", pg. 79

[6] To emphasize, note that the choice of comparison metric is an issue that arises *only* when we want to quantify the comparison. If we were only interested in *whether* Page A is better than Page B, then any metric (as long as it is monotonic in p_A and p_B) would do. The choice of metric becomes important when we are interested not just in whether Page A is better than Page B, but in addition, *by how much*.

To get an idea of what the metrics look like and to compare them, in Figure **??** we plot the lines defined in 2-dimensional probability space $\{(p_A, p_B)\}$ by a non-zero value for each metric. If the line corresponds to a acceptable value for the metric and the known success probabilities correspond to a point below and to the right of the line, then A is deemed better than B. **superimposed on pdf?**

Note that in reality we do not have a point (p_A, p_B) that corresponds to our knowledge of the success probabilities, instead we have the likelihood $f(p_A, p_B)$. Thus, we can ask for the *probability* that (p_A, p_B) lies below the contour line for a specific value, which is simply the volume of the likelihood $f(p_A, p_B)$ below the line, and is called the *credibility* of the result. We show how to calculate this in Section **??**

Thus, we use the data to infer not just *which treatment is better*, but in addition *by how much* and *how certain we are of this*.

Since the inferences about "better" are probabilistic for any finite data, the "more better" we want A to be —i.e. the larger the acceptable value for the difference— the less credible it is that A is better than B by that amount. In Figure **??**, the metric lines get squeezed into the bottom right corner as the metric value increases. As a one-dimensional example, if we obtained 60 Heads out of 100 trials, it is much less likely to have arisen from a heavily loaded coin (say $p = 0.80$) than from a moderately loaded coin ($p = 0.65$).

For concreteness, consider experimental data $(n_A, m_A) = (200, 40)$ and $(n_B, m_B) = (100, 15)$, and we will use the various metrics to illustrate results. The analysis we've described so far provides a *Metric vs. credibility* curve based on the experimental data, of the form in Figure 2 for the Probability Difference metric. The question then is how we can use this to make a recommendation

Figure 2: Prob. Difference vs. Credibility

about A vs. B. Recall that the business client has chosen a comparison metric and a minimum acceptable value (say 0.01) which will lead to a recommendation for a business action. From the above experimental curve, we or the client can certainly read off the credibility at a given acceptance value to obtain a statement like "0.02 difference has 70% credibility", but it is not yet clear how to compare this or similar statements to the client's single acceptance value. In our opinion, it is neither fair to place the onus on the business client of defining a 2-dimensional acceptance point "0.015 difference at 85% credibility", nor is it entirely mathematically consistent (See Section **??**.).

How do we consistently compare the single value for acceptance provided by the client to the credibility curves we've obtained? Note that the client's threshold M_{Acc} is effectively the acceptable value *at 100% credibility*, or the acceptable *expected* value, which we can now treat as a constant. On the other hand, from the data, for each value (of a given metric), we obtain a credibility, and we interpret the product of credibility and value as the *expected minimum value of the metric*. It is

Number of Trials	n
Number of Successes	m
Mean	$\mu^F = \frac{m}{n}$
Variance	$\mu^F \cdot (1 - \mu^F)$
Variance of the Mean	$\frac{\mu^F \cdot (1-\mu^F)}{n-1}$

Table 1: Descriptive Statistics of the Sample Results

straightforward to show that this expected minimum value has a maximum, as can be seen in Figure 3 where we've plotted the Expected Minimum Value of the Probability Lift vs. the Probability Lift itself. The question is then reduced to comparing the experimentally determined *maximum*

Figure 3: Expected Min(Lift) vs. Lift

expected minimum value to the client's acceptable value: if the *max-min* is lower then the variant is *not* better than the control.

In Section **??** we use the joint likelihood to compute the credibility of any value of one of the three comparison metrics. The credibility is calculated numerically, by dividing the integration domain into small quantiles of the likelihood (See Section **??** for details.).

Finally, for the same sample data as in the example above, we describe how to plot curves for the expected minimum value of the metric vs. either the credibility or the metric value itself, for any of the three metrics, e.g. Figure **??** Section **??**. From these curves the maximum can be easily read off and used to compare with the client's acceptable value and provide the client with a yes/no answer.

2 Summary of the Approach

2.1 The Data Model

As a result of the experiment, for each test-variant pair, we obtain

- n_A, m_A — the number, respectively, of trials and successes on Page A
- n_B, m_B — the number, respectively, of trials and successes on Page B

Table 1 summarizes some univariate descriptive statistics for the sample results of each page, which we will compare to those of the posterior distribution for the likelihood in Section 3.2

2.2 Summary of the Approach

We approach the problem as follows:

1. Given the input data (Section 2.1), calculate the Bayesian posterior distribution (the likelihood) of p_A, p_B as in Section 3.3.
2. Decide on a comparison metric M by which to *quantify the difference* between probabilities p_A and p_B. See Section 4.
3. Use the posterior distribution (computed in Step 1) to determine the difference vs. credibility curve for a given comparison metric, see Figure 2
4. From the above, calculate the expected minimum metric vs. credibility curve and the maximum thereof (see Figure 3), which can be compared to the client's acceptance value.

3 Bayesian Statistics

3.1 Computing the Probability Distribution for One Population

The probability of having m successes in n trials for a coin that has an intrinsic probability of success x is given by the Binomial distribution

$$P(m|x,n) = \binom{n}{m} x^m (1-x)^{(n-m)} \tag{1}$$

which is normalized over m

$$\sum_{m=0}^{n} P(m|x,n) = 1 \tag{2}$$

We wish to solve the reverse problem, namely, we wish to infer (probabilistically) the intrinsic property of the coin given an experimental outcome. For example, what is the likelihood that the coin is fair when we obtained 7 Tails out of 10 trials? In other words, given m, n, derive $f(x)$, the probability density of x.

We start by reminding the reader of Bayes' Theorem

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} \tag{3}$$

In our context, this is re-written as follows

$$f(x|n,m) = \frac{P(m|x,n) \cdot P(x)}{P(m)} \tag{4}$$

Consider the three terms on the right hand side

$P(x)$ Given that we have no prior knowledge of the true probability, $P(x)$ is assumed to be the uniform distribution i.e., $P(x) = 1$.

$P(m|x, n)$ This is the Binomial distribution Equation 1

$P(m)$ This is ill-defined but drops out when we normalize the probabilities to integrate to 1.

With this normalization the distribution $f(x|n, m)$ is the ***Beta distribution*** [7]

$$f(x|n, m) = \beta(x; m + 1, n - m + 1) = (n + 1)\binom{n}{m}x^m(1 - x)^{n-m} \tag{5}$$

3.2 Properties of the Beta Distribution

Some properties of the Beta distribution that are worth noting:

- The domain of the Beta distribution is the unit interval $I = [0, 1]$, and the Beta distribution is 0 at $x = 0, 1$ for all $n, m > 0$.
- As we have stated before, the distribution is normalized over I:

$$\int_0^1 dx \cdot f(x|n, m) = 1$$

- In the case where we have *no* data, i.e. $n = m = 0$, the posterior distribution is the uniform distribution

$$f(x|0, 0) = \beta(x|1, 1) = 1$$

- The mode, or the value of x at which the distribution is a maximum, is given by

$$\texttt{mode}(x) = \texttt{argmax}(\beta(x|m + 1, n - m + 1)) = \frac{m}{n}$$

  which is the experimental mean. Thus the experimental mean predicts the maximum likelihood.

- From the properties of the Beta distribution, the mean of the Bayesian distribution,

$$\mu = \int_0^1 dx \cdot x \cdot f(x|n, m) = \frac{m + 1}{n + 2} \tag{6}$$

[7] `https://en.wikipedia.org/wiki/Beta_distribution`

Note first that this is *not* equal to the mode, in fact the mean regresses from the mode towards 0.5, which, in the absence of other information is the mean probability of the outcome of a binary valued experiment. Note also the difference between $\mu^{Bayes} = \frac{m+1}{n+2}$ in Equation 6 and the mean of the experimental distribution $\mu^F = \frac{m}{n}$ in Table 1. This is because the Beta distribution adds 2 "pseudo-trials", one being a success and the other being a failure, to the observed trials. This is consistent with the Bayesian prior probability distribution $P(x) = 1$.

- The Beta distribution is not symmetric about the mean, except for the special case $a = b$, in which case $\mu = 0.5$:

$$\beta(x|a, b) = \beta(2 \cdot \mu - x|a, b) \iff a = b \tag{7}$$

- The variance of the distribution is

$$\sigma^2 = \frac{\mu \cdot (1 - \mu)}{n + 3} \tag{8}$$

which is different from the variance of the mean in Table 1 in two ways: the mean here is that of the posterior distribution, not the Frequentist mean as in the table, and the denominator includes the two extra "psuedo-trials".

- Finally, the Beta distribution satisfies an obvious multiplication law, which we've expressed in terms of $f(x)$

$$f(x|m, n) \cdot f(x|m_0, n_0) \propto f(x|n + n_0, m + m_0) \tag{9}$$

The implication of this is that the results of consecutive experiments are additive: Consider a situation in which the trial period of an experiment yields results (n_0, m_0) and the subsequent experiment yields (n, m). For consistency's sake, we require that the Bayesian posterior probability be determined by the cumulative results $(n + n_0, m + m_0)$. We see that this is true: the *prior* probability distribution $P(x)$ in Equation 4 arises as the posterior to results (n_0, m_0). Then $P(x) = f(x|n_0, m_0)$. Hence the probability distribution posterior to the subsequent experiment with results (n, m) is given by $P(m|x, n) \cdot f(x|n_0, m_0)$. The multiplication rule above, Equation 9 then implies that the consistency condition is satisfied.

3.3 Computing the Joint Probability Distribution for the Two Populations

Now that we have the individual PDFs for A, B

$$\begin{aligned} f_A(x) &= f(x|n_A, m_A) = \beta(x|m_A + 1, n_A - m_A + 1) \\ f_B(y) &= f(y|n_B, m_B) = \beta(y|m_B + 1, n_B - m_B + 1) \end{aligned} \tag{10}$$

we can compute the joint probability density $f(x, y) = P[p_A = x, p_B = y]$

$$f(x, y) = f_A(x) \cdot f_B(y) \tag{11}$$

where we use the fact that user behavior on page A is independent of user behavior on page B.

We now have the posterior probability refered to in Step 1 of Section 2.2

Metric	Section	Parameter	$\mathbf{M(x, y) > 0}$	**Boundary:** $\mathbf{y = m(x)}$
Difference	4.1	δ	$x - y - \delta$	$y = x - \delta$
Lift	4.2	λ	$x - y \cdot (1 + \lambda)$	$y = \frac{x}{1+\lambda}$
Odds Factor	4.3	ϕ	$O(x) - \phi \cdot O(y)$	$y = O^{-1}\left(\frac{O(x)}{\phi}\right)$

Table 2: Summary of Metrics

Row	n	$\mathbf{m_B}$	$\mathbf{m_A}$	y	x	Diff. δ	Lift λ	Odds F.ϕ
1				0.50	0.55	0.05	10%	1.2
2				0.05	0.10	0.05	100%	2.1
3				0.04	0.10	0.06	150%	2.7
4	$10k$	9	10	$0.9m$	$1.0m$	$0.1m$	11.1%	1.11
5	$10k$	$9 - 1 * \sigma = 6$	10	$0.6m$	$1.0m$	$0.4m$	67%	1.67
6				0.90	0.95	0.05	5.6%	2.1

Table 3: Metrics: Examples

4 Comparison Metrics

In this section, we elaborate on the metrics summarized in Table 2

4.1 Probability Difference

The intuition here is that we consider A to be better than B if the intrinsic probability p_A exceeds p_B by some constant δ. This metric has many drawbacks, but to illustrate one, consider rows 1 and 2 in Table 3: An increase in Row 1 from 0.50 to 0.55 feels intuitively very different from the increase in Row 2 from 0.05 to 0.10, yet the difference in both cases is the same 0.05.

4.2 Probability Lift

We could address the above drawback of the difference metric (Section 4.1) by measuring the increase in terms *relative* to Control. For example, one could want the Variant to result in 10% greater conversion ratio than the Control. The desired *lift* is often written as

$$\lambda = \frac{x - y}{y} \tag{12}$$

In Rows 1 and 2 in Table 3, the lift changes from 10% to 100% and captures our intuition about the "bigness" of the change. One problem with this metric is that often, a large lift is nothing more than a measure of the smallness of the effect in the Control group, and exposes the decision to the vagaries of an ill-chosen control group. A related problem with this metric is that it is *non-linearly*

sensitive to errors in the Control y. Consider the small drop in y from 0.05 to 0.04 between Rows 2 and 3, which causes a dramatic increase in the lift from 100% to 150%.

Such errors can easily arise due to poor statistics, even with a large number of trials. In Row 4, there were only 9 successes in the control group. The standard error in the number of successes is $\sqrt{9} = 3$, which translates into a (conservatively underestimated) 15% probability of getting 6 or fewer successes (Row 5). So without any likely *real* change in the property of the population, this "noise" causes the lift to jump by a factor of 6 from 11% to 67%, even though y has only decreased by a third!

Finally, neither of these two metrics captures the fact that it is just as important to reduce the failures as it is to increase the successes. Consider the difference between Rows 2 and 6:

1. In Row 2 an increase in probability from 0.05 to 0.10 corresponds to a lift of 100%
2. In Row 6 an increase in probability from 0.90 to 0.95 corresponds to a lift of only 5.6%

On the other hand, the jump from 0.90 to 0.95 in Row 6 means that we have done 50% as well as we could have, given that we could not go higher than 1. The jump from 0.05 to 0.10 in Row 2 means we did only 5.6% as well as we could have. Is there a metric that captures how much better we've done in comparison to how much better we *could have* done?

Consider also the following flaw with the above two metrics: From a methodological or product design standpoint, one should establish the acceptance value for the metric that triggers a business action *before* starting the experiment: first, so as to not get vested in the positive outcome of an experiment and second, because the business environment under which decisions are being made is mostly the same before and after the experiment. From this standpoint, both the difference and lift metrics fail. A reasonable acceptance value cannot be established *before* knowing what the control results are. For example, while a pre-established minimum lift of 100% is entirely reasonable if it turns out the control group's success rate is 0.05, a lift of 100% is absurd if the control group's success rate turns out to be 0.9.

For important but more abstract mathematical problems with the above two metrics, see the expanded version of this article.

4.3 Odds Factor and Related Metrics

A common measure of traffic to a website is the number P of page visits per session. This metric is useful since it is reasonable to assume that on average page visits is proportional to viewing opportunities to click or convert which in turn is linear in revenue. Suppose that the probability p that we've measured is the *page transition probability* for the website, the probability that the user visits another page on the website as opposed to leaving the website, timing out or otherwise ending

the session. Straightforward algebra shows that the average number of page visits per user-session is given by

$$P = p + p^2 + p^3 ... = \frac{p}{1-p} \tag{13}$$

This is nothing but the *odds ratio* or simply the odds corresponding to the probability. It overcomes the problems with the probability difference and the lift discussed above, and in addition, is both mathematically sound and familiar to non-technical people.

Definition 1 *For a probability p, the odds $O(p) = \frac{p}{1-p}$*

From the odds o, the probability can be recovered by

$$O^{-1}(o) = \frac{o}{1+o} \tag{14}$$

This is a metric familiar to gamblers. Why do gamblers think in terms of odds rather than probabilities? In a *pay-to-play* situation they intuitively understand that when evaluating a position they have to take into account the cost of failure $\$C$ as well as the benefits of success $\$B$. They are evaluating and optimizing the *benefit-to-cost-ratio*

$$\frac{\$B \cdot p}{\$C \cdot (1-p)}$$

Since both $\$B$ and $\$C$ are presumed independent of p, these drop out of consideration and one wants to increase the odds. So to compare the results of A and B one metric we will look at is

$$O(x) > \phi \cdot O(y) \tag{15}$$

This is interpreted as saying that the odds associated with A are at least $\phi\times$ the odds associated with B. We will consider minor modifications of this metric, such as the page views difference and the page views lift, in a future article.

Note in Table 3 that the odds factor is much more stable w.r.t. errors in y, in the sense that it behaves linearly. For example, between Rows 4 and 5 the control probability drops by a third, and the odds lift increases $1.5\times$, as opposed to the $6\times$ increase in the lift.

4.4 Comparison of the Metrics

In order to get a better feeling for the metrics, for each of them, Figure **??** shows their contours in 2D probability space for the comparison parameter values in Table **??**. We note that the ranges of both the difference metric and the lift are bounded and do not apply to all values of (x, y). The odds factor has no such limitations.

Metric	Section	Comparison Parameter	Value
Difference	4.1	δ	0.25
Lift	4.2	λ	1.0
Odds Factor	4.3	ϕ	2.0

Table 4: Values of Comparison Parameters for Fig. **??**

Figure 4: Probability Comparison Metrics on 2D Probability space

There are a few other things to note. When their threshold values are 0, the boundaries for all metrics collapse to the 45° diagonal line, and all the Bayesian credibilities will be the same. However, when the thresholds have non-zero values, the boundaries are *not* the same. As we can see in Fig.**??**, there can be large gaps between the different metric lines at all values of x. When the likelihood has signficiant support in either of those regions, the metrics will lead to contradictory decisions. **seems to call for a superimposed plot of pdf and metric contours?**

5 Comparing Pages A and B quantitatively

In order to quantify the comparison of Page A and Page B, we choose one of the metrics $M(x, y)$ from Table 2 discussed in Section 4 above. The metric may have a non-zero threshold as an argument. M is a mapping from the unit square $[0, 1] \times [0, 1]$ into $\mathcal{R}$. Given a comparison metric M, the *credibility* that $M > 0$ or the probability that Page A is better *in terms of metric* M than Page B is

$$\texttt{Credibility}[M > 0] = \int_{M(x,y)>0} f(x, y)\, dy\, dx \tag{16}$$

Most reasonable choices for M are such that we can rewrite Equation **??** as

$$\int_{x=0}^{x=1} f_A(x) \left(\int_{y=0}^{y=m(x)} f_B(y)\, dy \right) dx \tag{17}$$

where $m(x)$ is the solution, for y, of $M(x, y) = 0$. (See the last column of Table 2.)

We will evaluate the double integral in 2 steps. In the first step, we note that the term within the parentheses has a closed form solution $\beta_{cf}(m(x); m_B + 1, n_B - m_B + 1)$, where β_{cf} is the cumulative function of the Beta distribution. This allows us to rewrite Equation **??** as

$$\texttt{Credibility} = \int_0^1 dx \cdot \beta(x, m_A + 1, n_A - m_A + 1)\beta_{cf}(m(x), m_B + 1, n_B - m_B + 1) \tag{18}$$

which is then evaluated numerically.

5.1 Implementation Detail: Integrating the Product of the Beta Distributions

The Python package `scipy.stats` includes the β distribution `beta.pdf(p|a,b)` and its cumulative function `beta.cdf(p|a,b)`

$$\beta_{cf}(x|a,b) = \int_0^{x'} dx' \cdot \beta(x'|a,b) \tag{19}$$

For numerical plotting routines or for numerical integration one discretizes the domain x into a set of points at which to evaluate the beta distribution or its cumulative function. For any number of points which break up the domain into uniform intervals, and for any given tolerance for error, there exists a number of trials n which is large enough that the effective support of the Beta distribution is smaller than an interval, resulting in graphical or numerical integration errors which exceed the tolerance. The graph will be either flat or contain an arbitrary valued spike and the integral will jump from 0 to some absurdly large value.

So what can we do? Very usefully, the quantile function or the inverse of the cumulative function `beta.ppf(percentile|a,b)` —which allows one to calculate the values of p at which a given percentile of the distribution occurs— is also implemented as a part of `scipy.stats`. Intuitively, what we want to do is to integrate or plot over a subset of the domain with some large proportion of the support. Assuming that we can tolerate an error of 0.2%, we choose a subset with 99.8% of the support. Then we can use `beta.ppf` to calculate the minimum and maximum values of the discrete set as

$$\begin{aligned} x_{0.1\%} &= \texttt{beta.ppf}(0.001|\texttt{m}+1, \texttt{n}-\texttt{m}+1) \\ x_{99.9\%} &= \texttt{beta.ppf}(0.999|\texttt{m}+1, \texttt{n}-\texttt{m}+1) \end{aligned} \tag{20}$$

One is again tempted to break up the subset $[x_{0.1\%}, x_{99.9\%}]$ into uniform intervals, but in fact the thing to do is to break it up so that the intervals are smaller where the distribution is larger and vice versa. This corresponds to finding the $x-$values for *uniformly distributed quantiles* —we are effectively discretizing the *range* of the cumulative function uniformly, *not* the domain of the PDF:

$$\texttt{Array(x)} = \texttt{beta.ppf}(\texttt{Array}([0.001, 0.999, \texttt{step} = 0.001])|\texttt{m}+1, \texttt{n}-\texttt{m}+1)$$

Since "(99.8% of) everything" is happening in this range of x values, using this array for numerical integration limits the errors.

5.2 Choosing an array of metric values

Due to considerations similar to the ones discussed above for integrating the probability distribution, one needs a "dynamic" way of determining a "good" range of metric values at which to evaluate the credibilities. Most low or high values of the metric will have credibilities of 1 or 0

respectively[8], and will vary between those extremes only for values of the metric where the metric line crosses the support of the probability distribution. No matter how fine-grained the static metric array is, there will be some narrow probability distribution which causes the credibility to abruptly jump from 0 to 1. Our approach to solving this is as follows:

1. Calculate the mean and variance of the Bayesian distribution for the probabilities $x = p_A, y = p_B$.
2. Calculate approximations to the mean and the variance of the metric $M(x, y)$ using the results derived in `mean_and_variance_of_function.pdf`.
3. Use these values in a normal approximation to the distribution of the metric to calculate the array of metric values corresponding to the percentiles

$$\begin{aligned} \texttt{Array}(\texttt{min}(\texttt{M})) =& \texttt{norm.ppf}(\texttt{Array}([0.01, 0.99, \texttt{step} = 0.01]) \\ & |\texttt{loc} = \texttt{mean}, \texttt{scale} = \texttt{sqrt}(\texttt{var})) \end{aligned} \tag{21}$$

5.3 Maximum Expected Minimum Value

For each value of M in the array constructed above, we can use Equation **??** to calculate the credibility

$$\texttt{Array}(Cred) = \texttt{credibility}(\texttt{Array}(min(M))) \tag{22}$$

These arrays can be used to plot the metric vs. credibility curves. Note that this is the credibility of the minimum value of the comparison metric, i.e. the probability that the actual value of the metric is larger than this minimum. Is this not sufficient to compare to the client's acceptance value M_{Acc}? Recall that M_{Acc} is implicitly the value that the client will accept (for the test to trigger a recommendation for A over B) at 100% credibility. From the perspective of the expected outcome, this is equivalent to their accepting a higher metric value at lower credibility, e.g. $1.11 * M_{Acc}$ at only 90% credibility since the expectation values are the same $M_{Acc} * 1.0 = 1.11 * M_{Acc} * 0.90$. So the client's expected acceptance value defines a curve in metric vs. credibility space

$$M * Cred(M) = M_{Acc} \tag{23}$$

This curve is a hyperbola and does *not* belong to the family of (inverse) sigmoidal curves as do the metric vs. credibility curves, e.g. Figure 2. In general, the client acceptance curve will intersect the metric vs. credibility curve in 0 or 2 points[9].

Consider the situation where the client has been asked to provide a 2-dimensional decision point $(Cred_{Acc}, M_{Acc})$. This defines an acceptance curve of points which are equivalent to each other

[8]For example, when the metric value is small and such that the metric line is to the left and above the support of the probability distribution, the credibility or volume under the curve will be 1.

[9]Ignoring the pathological situation where they are tangent.

from an expected value perspective:

$$M * Cred(M) = M_{Acc} * Cred_{Acc} \tag{24}$$

If there are no intersections the recommendation is unambiguous, A is not better than B. However, if there are two intersections, the recommendation is ambiguous since it implies that there are points on the acceptance curve below the metric vs. credibility curve that trigger a "Yes" recommendation but that there are equivalent points on the acceptance curve that are above the metric vs. credibility curve that do not trigger a recommendation. This is the ambiguity in the two-dimensional decision point approach we referred to earlier.

So we require from the client a single expected acceptance value M_{Acc} which defines a curve Equation **??**. From the credibility array Equation **??** we can construct the array of expected minimum metric values since

$$Expected(min(M)) = Cred(M) * min(M)$$

One can show that the $Expected(min(M))$ has a maximum which we denote by $\texttt{MaxExpMin}(M)$. This can be found from the above array and can be visualized e.g. in Figure 3 where the $Expected(min(M))$ of the probability Lift has been plotted against the va;lue of the lift itself. To make a recommendation, we proceed as follows.

5.4 Recommendation

The maximum expected minimum value determined above from the data is compared to the *Acceptance* value M_{Acc} for the metric M established by the business client:

Definition 2 *If* $\texttt{MaxExpMin}(M) \leq M_{Acc}$ then A is not better than B

Like most scientific tests, strictly speaking it is only good for ruling out the null hypothesis. However, we can recommend

Definition 3 *If* $\texttt{MaxExpMin}(M) > M_{Acc}$ then A is better than B

6 Sample results for Bayesian Credibility

For test results with $(n_A, m_A, n_B, m_B) = (200, 40, 100, 15)$, we plot the credibility as a function of the importance levels for the three metrics proposed above. **show all three curves for one metric only as example?**

6.1 Probability Difference

From Figure2 the user can read-off the probability difference that corresponds to a chosen credibility, or vice versa. We see that as expected the credibility drops as the difference between p_A and p_B increases. As described in the Introduction Section 1, the business client or analyst can compare the position of the metric-credibility curve in the figure to the decision point (threshold metric, credibility) and decide whether to proceed with the business action.

6.2 Probability Lift

From Figure 3 the user can read-off the maximum expected minimum lift which can be compared to $Lift_{Acc}$.

6.3 Odds Factor

Figure 5: Expected Min(Odds Factor) vs. Credibility

From Figure **??** the user can read-off the maximum expected minimum odds factor[10].

7 Conclusion

For both the test and control populations of a binary outcome A/B test the number of trials and successes are counted. We've taken the Bayesian approach to calculate the posterior joint likelihood for the intrinsic success probabilities of the two populations from the experimental data. Separately, and possibly before the experiment is started, based on the business outcome desired a comparison metric is chosen, and an *acceptance value* M_{Acc} which will trigger a recommendation. For the chosen metric, a metric vs. credibility curve is calculated from the joint probability distribution, and from this in turn a plot of the expected minimum metric vs. the credibility. The maximum of this is found, the `MaxExpMin`(M). A business action is recommended based on comparing `MaxExpMin`(M) to M_{Acc}.

8 Appendix

In this section we will briefly outline the Frequentist approach and then compare it to the Bayesian approach we have taken in this paper.

[10]As an aside, we point out that, due to the symmetries of the odds and those of the Beta distribution, the Odds Factor vs. credibility curve is invariant under the discrete symmetry

$$(n_A, m_A, n_B, m_B) = (n_B, n_B - m_B, n_A, n_A - m_A)$$

.

8.1 The Standard Frequentist Approach

The material in this section is well-known and is included for completeness and comparison. For an utterly convincing argument that demonstrates the superiority of the Bayesian approach without assuming any *a priori* knowledge of statistics, please take the time to read `https://xkcd.com/1132/`.

8.1.1 Single Variant

Consider the situation for one variant as described earlier in Section 3.1, with n trials indexed by i and the corresponding outcomes $x_i \in \{0, 1\}$. Let there be m successes (1) and $n - m$ failures (0). Then the mean or expectation of $\{x_i\}$ is

$$\langle x \rangle = \mu = \frac{1}{n} \sum_i x_i = \frac{m}{n}$$

and the variance of the distribution is

$$var(x) = \sigma_x^2 = \mu \cdot (1 - \mu)$$

where σ_x is the standard deviation.

Now, we are primarily interested in the variance or the Standard Error in the estimated mean. From the definition of variance, it is fairly easy to show that the variance of the mean is

$$var(\mu) = \sigma_F^2 = var(\langle x \rangle) = \frac{var(x)}{n} = \frac{\mu \cdot (1 - \mu)}{n}$$

One then uses the above parameters to calculate the z-score corresponding to some proposition about μ, assumes that μ is normally distributed and uses the cumulative function of the Normal distribution to calculate the percentiles or p-value from the z-score.

8.1.2 The confidence that A is better than B

Suppose we were interested in the proposition that $M(x, y, \delta) > 0$, where x, y are the probabilities associated with variants A and B respectively and $M(x, y)$ is one of the comparison metrics defined earlier with minimum acceptable value δ. If we had estimates for $\langle M \rangle$ and $SE(M) = \sqrt{var(M)}$ then we could calculate the z-statistic for the proposition via

$$z(M) = \frac{\langle M \rangle}{SE(M)} \tag{25}$$

and proceed to calculate the parametric confidence level or p-value. Since we are only interested in whether A is better than B, the (1-sided) proposition we want to test is whether the probability associated with variant A is greater than the probability associated with variant B, i.e., $M(x, y) = x - y > 0$. The proposition is

$$M(x, y) = x - y > 0 \tag{26}$$

from which

$$\begin{aligned} \langle M \rangle &= \langle x \rangle - \langle y \rangle \\ var(\langle M \rangle) &= var(\langle x \rangle) + var(\langle y \rangle) \\ z(x > y) &= \frac{\langle x \rangle - \langle y \rangle}{var(\langle x \rangle) + var(\langle y \rangle)} \end{aligned} \tag{27}$$

8.2 Reasons to Prefer Bayesian to Frequentist

Our reasons to prefer the Bayesian approach to the Frequentist are listed below:

1. The Frequentist approach involves a parametric approximation usually based on the normal distribution, which is not defined on probability space and is notoriously inaccurate for situations with low numbers of trials as well as those with probabilities near 0 or 1.

   The Bayesian approach doesn't assume normality, it derives the form of the posterior distribution.

2. In the Frequentist approach, the results of the parameteric calculations depend on the the metric used for comparing the two probabilities p_A and p_B. In the first place normality is not preserved under nonlinear transformations, e.g. if x is normally distributed, e^x is not. In the second place, the calculation of the descriptive statistics for non-linear metrics is somewhat non-trivial and involves an approximation[11]. Furthermore, the actual values of the confidence levels depend on the algebraic form of the metric. For example,

   $$\begin{aligned} &P[\frac{1-y}{1-x} - (1 - \lambda_0) > 0] \quad \text{and} \\ &P[(1 + \lambda_0) \cdot x - y - \lambda_0 > 0] \end{aligned} \tag{28}$$

   are the probabilities of logically equivalent propositions but the outlined approximations to calculating them lead to different values of the confidence.

   The Bayesian approach (at least in the case of the Binary tests considered here, for which the posterior distributions can be constructed and the required integrations carried out numerically) can be applied to any comparison metric on 2D probability space, is well-defined and does not need to be approximated.

[11] See `mean_variance_of_function.pdf`.

3. The Frequentist/parametric approach breaks down when either trials or successes are small in number and one has to take care with the use of the standard errors.

   When the numbers are small, the Bayesian distribution has a large variance that reflects our uncertainty in the intrinsic probability and it can be applied without treating the small number situation as a special case.

For a discussion of the above reasons, for details on calculating the Frequentist confidence levels for the different metrics described here in Section 4 and for limits on the domain of validity for which Frequentist calculations approximate the Bayesian ones, see `frequentist_vs_bayesian.pdf`.